Détection de fraudes et loi de probabilité de Newcomb-Benford

FERNANDEZ Christelle PONCHEELE Clément EL KAÏM Laura
Encadré par M.DUCHARME

5 *mars* 2021

RESUME DU PROJET EN QLQ LIGNES

REMERCIEMENTS

Table des matières

Introduction.

La fraude est une pratique répandue dans de nombreux domaines comme par exemple la finance, le secteur social ou médical. Il peut être tentant pour un être humain ou une société de tricher si cela peut impliquer pour lui une position plus confortable dans la société, telle qu'une réduction de charges, ou même un avantage sur un de ses concurrents. Il semblerait donc logique que des personnes cherchent à déceler ces fraudes.

Les données transmises par un individu ou un organisme peuvent faire l'objet de modifications, c'est de ce type de fraudes auquel nous nous intéresserons ici, et plus particulièrement la modification du premier chiffre significatif (le premier chiffre d"un nombre qui n'est pas un zéro) de nombres pris dans un certain ensemble de données.

De telles modifications entraînent un changement de la répartition des chiffres présents naturellement [1].Si nous connaissons la répartition des chiffres présentés dans un ensemble de données arbitraires, il est donc techniquement possible de savoir si un nombre a été modifié ou non.

Il nous vient donc les questions suivantes : *Qu'elle est cette répartition ? Est-il possible de la connaître et si oui, dans quels cas ?*

De manière intuitive nous pourrions penser que les nombres sont répartis de manière uniforme. Qu'en est-il vraiment ?

La première partie de notre projet consistera à **répondre à ces questions**, nous nous appuierons sur les travaux de Simon Newcom et Frank Benford, qui ont théorisé la **loi de Newcomb-Benford**, plus communément appelée loi de Benford. Cette loi nous dit que, dans une liste de données dites naturelles, la probabilité d'avoir le chiffre i comme premier chiffre significatif est de $log_{10}(1 + \frac{1}{i})$.

Par exemple, le chiffre 1 en tant que premier chiffre significatif serait présent à hauteur de 30% alors que le 9 à seulement $4,6\%$.

Dans la suite **nous mettrons en œuvre une série d'expérimentation** pour constater ou non la véracité de cette loi, pour ce faire dans un premier temps nous récolterons des nombres pris dans des milieux sensés satisfaire la loi de Newcomb-Benford et observerons la répartition du premier chiffre significatif. Puis nous répliquerons une version simplifiée de l'expérience de Hill (1988), qui consiste à observer la répartition du premier chiffre significatif d'une liste de nombre donnée au hasard par des êtres humains, en l'occurrence ses élèves.

Cette expérience est à la base des méthodes de détection de fraudes par la loi de Newcomb-Benford. Si un fraudeur modifie un jeu de données, ce jeu est donc influencé par la pensée humaine, il ne suit donc plus la loi de Newcomb-Benford. Pour detecter la fraude il suffit donc de comparer les premiers chiffres significatifs. Cependant ces comparaisons doivent se faire de manière rigoureuses et scientifiques. Pour cela il existe des test statistiques, dont le plus connu, le test du χ^2, ou bien celui de Ducharme et collab. (2020).

Il nous vient donc les questions suivantes : *Ces test sont-ils fiable ? Existe-t-il un test significativement meilleur que les autres ? Vont-ils dans le même sens ? Et sinon que faire ?*

La réponse à ces question constituera donc la deuxième partie de ce projet, pour ce faire nous mettrons en oeuvre différents tests sur des jeux de données comme la fiscalité italienne.

1. Les données dites naturelles sont celles qui qui n'ont pas été influencé par la pensée de l'homme. .

Naissance de la loi de Newcomb-Benford.

Il serait tentant de penser que les nombres sont répartis de manière uniforme, cela viendrait du biais d'équiprobabilité [2]. Ce dernier consiste à "penser qu'en l'absence d'information, tous les cas ont la même probabilité de se produire et que le hasard implique nécessairement l'uniformité".

Néanmoins cette hypothèse sera contredite une première fois par l'astronome, mathématicien, économiste et statisticien canadien Simon Newcomb. Ce dernier fournira en 1881 une première approche au principe statistique, qui se fera injustement appeler *Loi de Benford*. Celui-ci remarquera que les premières pages des tables logarithmiques sont plus utilisées que les pages suivantes. Il publiera sa découverte dans un article de l'*"American Journal of Mathematics"*.

Cette découverte mise de côté pendant plusieurs années, ce n'est qu'en 1938 que l'ingénieur et physicien américain Frank Benford arrivera au même résultat après avoir répertorié des dizaines de milliers de données. Celui-ci pensera être le premier à l'initiative de cette loi, et c'est pour cette raison que la *loi de Newcomb-Benford* se fera plus généralement appelée *loi de Benford*.

Cette loi nous dit que, dans une liste de données arbitraires, la probabilité d'avoir le chiffre i comme premier chiffre significatif est de $log_{10}(1 + \frac{1}{i})$.

PCS	1	2	3	4	5	6	7	8	9
Benford	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Tableau 1 : Répartition du premier chiffre significatif selon la loi de Newcomb-Benford.

Nous retrouvons cette loi dans énormément de domaines comme les mathématiques, l'environnement, la finance, la physique, etc, plus précisément sur des données telles que la longueur des fleuves, la population des villes dans un pays, des déclarations de revenus, etc.
Notons cependant qu'il existe des cas où les données ne suivent pas cette loi, notamment des données dites non naturelles qui seraient influencé par la pensée humaine (nombres premiers, nombres générés par des humains, etc).

2. Défini en 1985 par Marie-Paule Lecoutre (*source*).

Expérimentation sur différents jeux de données.

Après avoir pris connaissance de la **loi de Newcomb-Benford**, il serait intéressant de la mettre en pratique sur différents jeux de données.

Données naturelles.

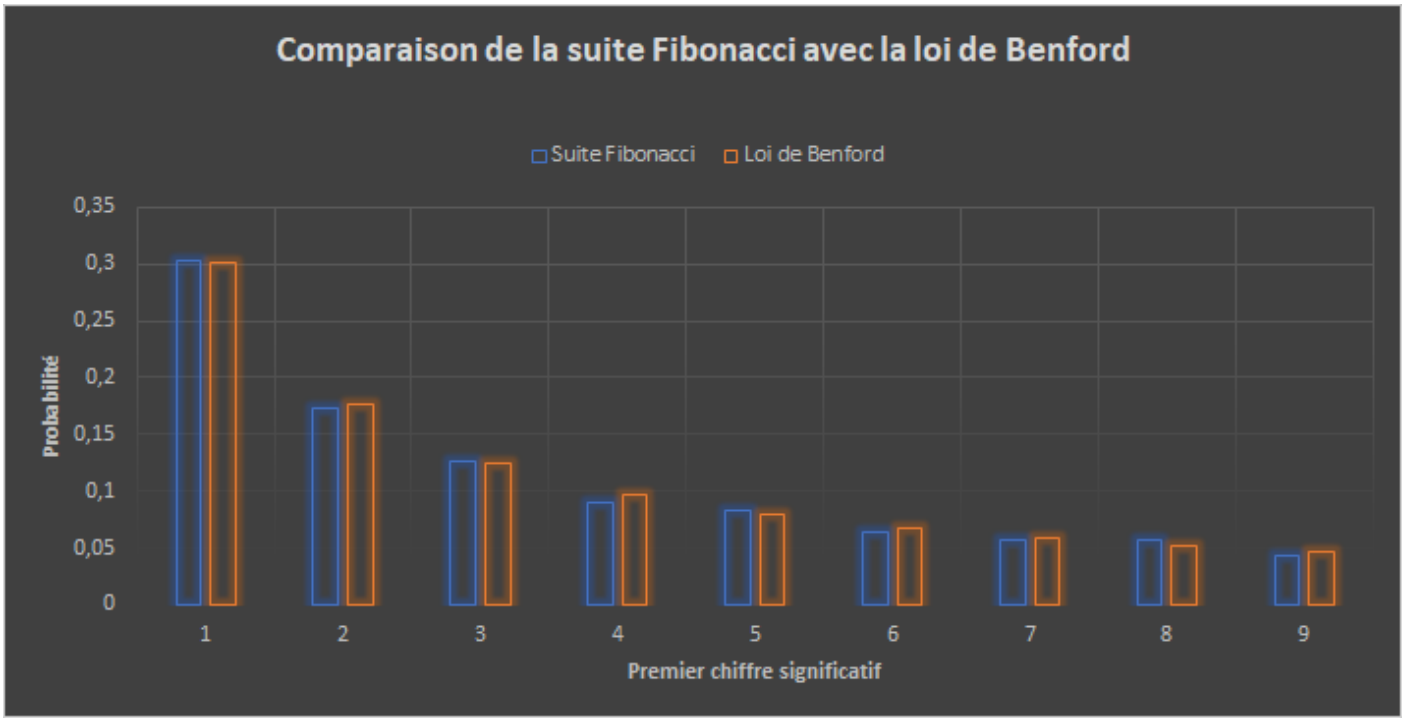
Intéressons-nous dans un premier temps à la suite de Fibonacci.

Cette suite est une suite d'entiers dans laquelle chaque terme est la somme des deux termes qui le précèdent. Sa formulation est la suivante :

$$F_0 = 0,\ F_1 = 1,\ \text{et}\ \forall\, n \geq 2,\ F_n = F_{n-1} + F_{n-2}.$$

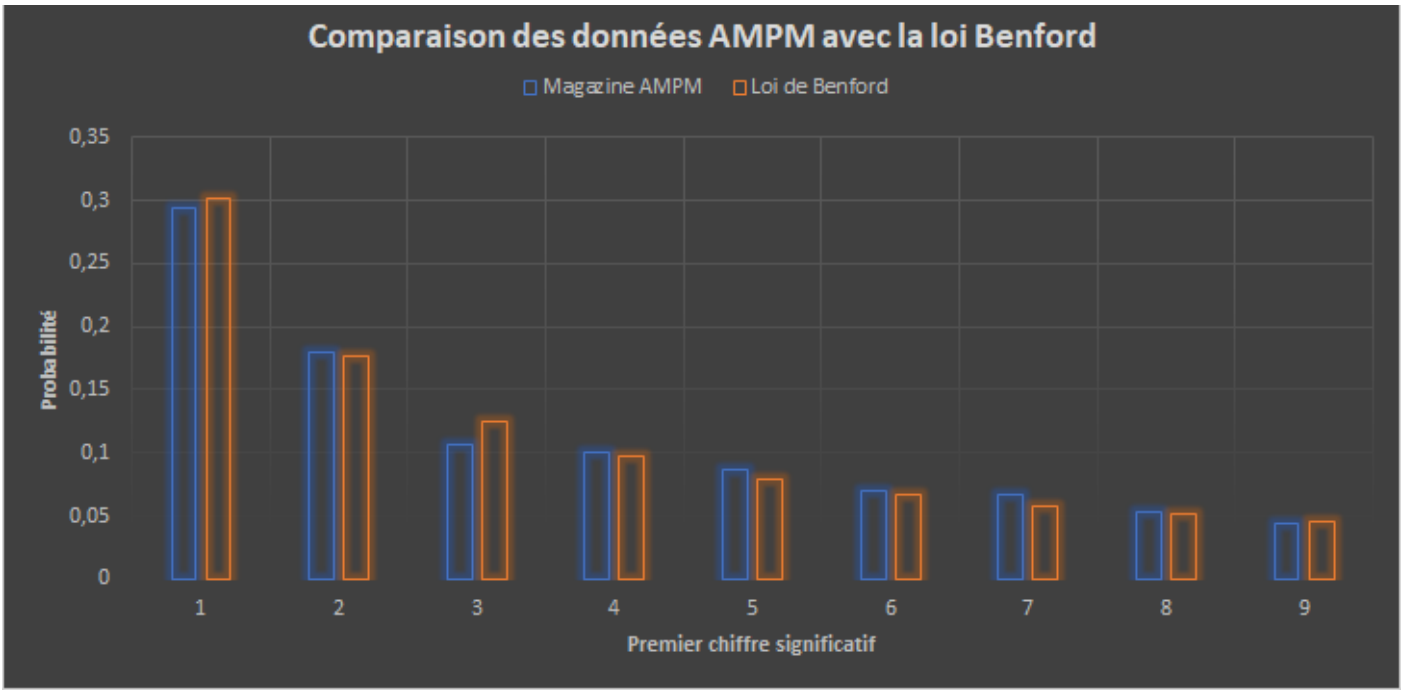
Nous commençons par recueillir les 300 premiers termes de la suite de Fibonacci, pour extraire le premier chiffre significatif de chacun de ces nombres.

Par la suite nous calculons la répartition de chaque chiffre significatif et obtenons l'histogramme suivant :



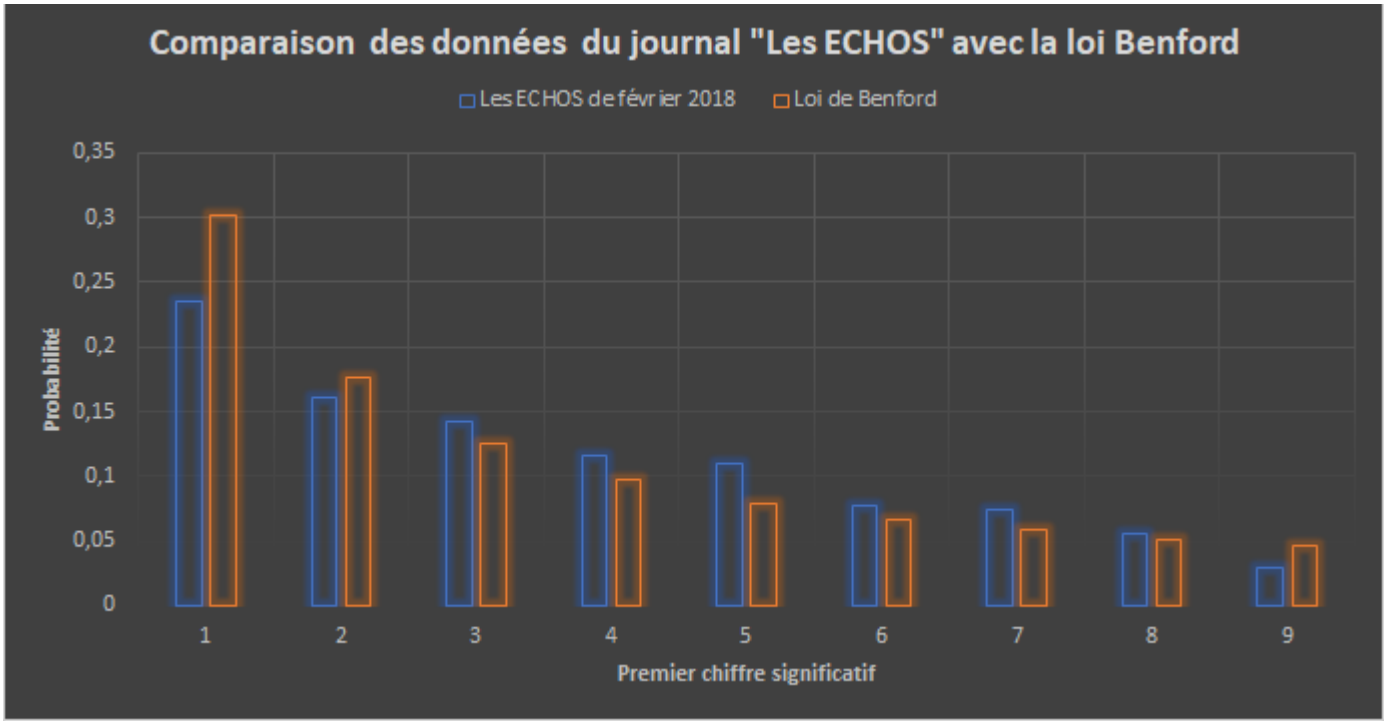
Il semblerait que la répartition des chiffres significatifs des 300 premiers nombres de la suite de Fibonacci suive la **loi de Newcomb-Benford**.

Dans un second temps, nous relevons les prix présents dans un magazine de mobilier de la marque *AMPM*, ainsi que tous les nombres répertoriés dans un journal *Les ECHOS*. Nous récoltons environ 300 nombres par magazine et, de la même façon qu'énoncé précédemment, calculons la répartition des chiffres significatifs de ces nombres.



La répartition des chiffres significatifs des prix du magazine *AMPM* parait fortement similaire à celle de la **loi de Newcomb-Benford**. Nous constatons tout de même une légère différence pour le chiffre 3.

Observons maintenant la répartition des données issues du journal *Les ECHOS*.



Nous remarquons ici la même tendance décroissante. Cependant les proportions des chiffres significatifs entre les données du journal et celles de la **loi de Newcomb-Benford** sont relativement différentes.

Pas Benford

Tests

Bibliographie